GMM-BASED SALIENCY AGGREGATION FOR CALIBRATION-FREE GAZE ESTIMATION

Jinsoo Choi, Byungtae Ahn, Jaesik Park, and In So Kweon

Korea Advanced Institute of Science and Technology,
Daejeon, Republic of Korea

ABSTRACT

A typical gaze estimator needs an explicit personal calibration stage with many discrete fixation points. This limitation can be resolved by mapping multiple eye images and corresponding saliency maps of a video clip during an implicit calibration stage. Compared to previous calibration-free methods, our approach clusters eye images by using Gaussian Mixture Model (GMM) in order to increase calibration accuracy and reduce training redundancy. Eye feature vectors representing eye images undergo soft clustering with GMM as well as the corresponding saliency maps for aggregation. The GMM based soft-clustering boosts the accuracy of Gaussian process regression which maps between eye feature vectors and gaze directions given this constructed data. The experimental results show an increase in gaze estimation accuracy compared to previous works on calibration-free method.

Index Terms— Gaze estimation, Saliency, Gaussian mixture model, Gaussian process regression

1. INTRODUCTION

Gaze estimation has gathered much attention due to its various applications such as gaze sensing for marketing analysis, interactive displays, and other human-computer interaction applications [1]. Gaze sensing can achieve sufficient human attention tracking for the mentioned applications.

There are two main classes for gaze sensing, namely model-based and appearance-based methods. Model-based methods construct 3D eye models along with iris contour to determine gaze directions. They are mainly concerned with locating the pupil center and thus require high resolution IR imaging and *special camera settings* which capture eyes very closely. The difficulties of model-based methods are that the geometry-based computations require sophisticated and extensive calibration. On the other hand, appearance-based methods are not computationally expensive and are robust in *ordinary capturing conditions*; a user looks at a monitor and a webcam is mounted on the monitor. Appearance-based methods can be implemented using a single camera device. However, previous appearance-based gaze tracking methods require an explicit personal calibration process. This calibration stage usually involves many discrete training fixation points leading to an unnatural interaction [2, 3]. In a display environment, the person must stare at numerous discrete points on the screen while a camera captures the person's eye images. Mapping the points to the eye images constructs the gaze estimator. This is a tedious and time consuming process and is one of the limitations of appearance-based methods.

Therefore, many gaze estimation approaches attempt to remove the need for explicit personal calibration. Chen *et al.* [4] combined image saliency with a 3D eye model. But since it is a model-based approach, it requires special camera settings. Sugano *et al.* [5] proposed a calibration-free method in a computer using environment. It basically assumes that the user looks at the position of the mouse during a clicking action. Yamazoe *et al.* [6] used an eyeball model and fit the model to the user's eye appearance for automatic calibration. However, It also requires special camera settings. Alnajar *et al.* [7] proposed calibration by matching gaze patterns of other humans. But this approach also needs specific prior datasets concerning numerous people and thus time consuming. Nguyen *et al.* [8] proposed a Bayesian approach to estimate gaze directions. It requires tunnable parameters for better gaze estimation accuracy. Sugano *et al.* [9] constructed a gaze estimator using eye images captured from a user watching a video clip and the saliency maps of each frame. A similar method [10] was done in addition to a feedback loop to optimize constituent saliency channel weights. They aggregate saliency maps and connect corresponding eye appearances to build a calibration-free gaze estimation method. However, saliency aggregation of this method is based on hard clustering built upon heuristic measurements.

Our approach follows the gaze estimation pipeline by Sugano *et al.* [9, 10]. In this pipeline, saliency aggregation is essential to improve gaze estimation quality and reduce data redundancy and computational cost. Compared to these literatures [9, 10], our work puts an emphasis on saliency aggregation and focuses on constructing a probabilistic framework rather than relying on heuristic measurements. We do this by constructing the GMM using an Expectation Maximization(EM) algorithm in the eye feature space.

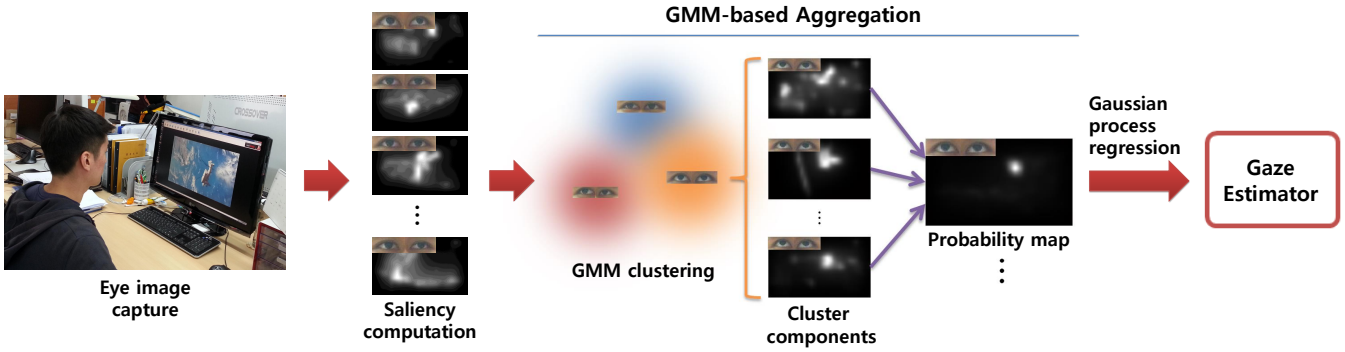


Fig. 1. An overview of our gaze estimation system.

2. THE GAZE ESTIMATION SYSTEM

Our system is similar as the calibration-free gaze estimation pipeline proposed by Sugano *et al.* [9]. The experimental framework involves the user to watch a video clip. While the user watches the video, a camera in front captures the eye image for all video frames. Also, the saliency maps are extracted using Graph-based Visual Saliency (GBVS) [11] algorithm for all video frames. This enables us to construct eye images and corresponding saliency pairs. With this data set, we can adaptively cluster the eye images (represented by eye feature vectors). Clustering is done by an EM algorithm to adaptively create Gaussian mixtures. According to the constructed GMM we can produce a weighted average eye feature vector and saliency map pair of all clusters. The weighted average of saliency maps can be treated as a probability map of gaze points. With this information, Gaussian process regression(GP) learns the eye feature and gaze point mappings to produce the gaze estimator. The gaze estimation system is shown in Fig. 1.

3. SALIENCY AGGREGATION

Saliency maps give a rough cue of where the person is looking. Thus, it carries valuable prior knowledge of the gaze but individually does not always give reliable true gaze points. The user may occasionally gaze at a non-salient point in the video frame etc. Therefore in order to increase reliability of gaze point estimation, saliency maps with similar eye images must be aggregated to produce a 'salient peak' around the true gaze. Like this, we can increase the gaze estimation accuracy as well as reduce the redundancy of training data.

In subsections 3.1 and 3.2, we give a theoretical explanation on aggregation by GMM clusters and resulting datasets. In subsection 3.3, we explain in detail how this is implemented using the EM algorithm.

3.1. Aggregation by GMM clusters

In a mixture of Gaussian distributions, an observed eye feature vector $e^{(i)}$ has a corresponding latent variable $z^{(i)}$ that specifies the mixture component that $e^{(i)}$ belongs to. Given $z^{(i)} = j$, we can say that $e^{(i)}$ is simulated from a normal distribution with μ_j and Σ_j as parameters. This is mathematically shown in (1):

$$(e^{(i)} \mid z^{(i)} = j) \sim \mathcal{N}(\mu_j, \Sigma_j). \tag{1}$$

Then by (1), we can say that:

$$P(e^{(i)} \mid z^{(i)} = j) = \mathcal{N}(e^{(i)} \mid \mu_j, \Sigma_j). \tag{2}$$

With this information, we can construct a weighted average of eye features and saliency maps for each cluster:

$$\overline{e}_j = \frac{\sum_i^N P(e^{(i)} \mid z^{(i)} = j)(e^{(i)})}{\sum_i^N P(e^{(i)} \mid z^{(i)} = j)}, \tag{3}$$

$$\overline{p}_j = \frac{\sum_i^N P(e^{(i)} \mid z^{(i)} = j)(s_i - s_{all})}{\sum_i^N P(e^{(i)} \mid z^{(i)} = j)}, \tag{4}$$

where N denotes the number of all training pairs, s_i is i-th saliency map and s_{all} denotes average of all saliency maps.

Notice how $P(e^{(i)} \mid z^{(i)} = j)$ from (2) can be used as weights for the weighted average of saliency maps. This makes sense in that it follows the same weighting scheme as the weighted average of eye feature vectors. Now we have the dataset:

$$D_p = \{(\overline{p}_1, \overline{e}_1), ..., (\overline{p}_M, \overline{e}_M)\}, \tag{5}$$

where $\overline{p}_i$ from (4) can be used as a probability distribution of gaze positions.

3.2. Inferring gaze points from probability maps

Originally, Gaussian regression (GP) estimates $P(g^* \mid e^*, D_g)$ where D_g denotes the dataset containing gaze positions instead of probability maps. We only have D_p from (5) and therefore we must somehow infer gaze positions from probability maps. In theory, a reformulation of GP can be established in order to predict $P(g^* \mid e^*, D_p)$ by:

$$P(g^* \mid e^*, D_p) = \sum_{g_1} \cdots \sum_{g_M} P(g^* \mid e^*, D_g) P(D_g \mid D_p). \tag{6}$$

However, the summation is computationally expensive, so we approximate D_g by taking the maximum posterior points from $P(g^* \mid e^*, D_p)$ as gaze points.

3.3. Adaptive clustering with EM algorithm

Here we specify the implementation details of aggregation by GMM clustering by the EM algorithm. We formulate an adaptive EM algorithm fit for our use of clustering eye feature vectors. The following is the algorithm:

(E-Step) For each i, j set:

$$w_j^{(i)} := P(e^{(i)} \mid z^{(i)} = j; \mu_j, \Sigma_j). \tag{7}$$

(M-Step) Update parameters if $w_j^{(i)} > \tau$:

$$\mu_j := \frac{\sum_{i=1}^{N} w_j^{(i)} e^{(i)}}{\sum_{i=1}^{N} w_j^{(i)}}, \tag{8}$$

$$\Sigma_j := \frac{\sum_{i=1}^{N} w_j^{(i)} (e^{(i)} - \mu_j)(e^{(i)} - \mu_j)^T}{\sum_{i=1}^{N'} w_j^{(i)}}, \tag{9}$$

where w_j is the weight for the jth cluster, and τ denotes the weight threshold.

In the E-step, we calculate the weights of $e^{(i)}$'s as shown in (7). Since we do not know how many clusters there should be, we assign $e^{(i)}$ to cluster j whenever $w_j^{(i)} > \tau$. Otherwise, clusters are adaptively created. In the M-step, equations (8) and (9) uses (7) as weights for updating parameters.

3.4. Comparison with other approaches

Our work emphasizes the saliency aggregation part with a theoretical approach. Sugano *et al.* [9] applies hard clustering of the eyes by a heuristic similarity measuring method. Specifically, an eye feature belongs to a cluster if it is analogous to the mean of the cluster. If it is not, the eye feature makes another cluster. The fixed threshold is used for the affinity decision. The next work by Sugano *et al.* [10] implements a feedback loop to refine saliency maps but is essentially the same as [9]. Therefore, [9, 10] suffers from bias and relatively low accuracy. Our method offers a natural way of soft clustering as shown in Fig. 2(b) for both eye feature vectors and saliency maps.

Intuitively, it makes sense to cluster eye features with a probability distribution (Gaussian in our approach). Within clusters there will be subtle differences in features that each correspond to different gaze points and thus, must be probabilistically modeled to compensate for it. Since gaze estimation is very sensitive to training inputs, it is essential to incorporate this operation. Notice from Fig. 2 that [9, 10] imposes hard cluster assignments for eye feature vectors. On the other hand, our method has soft assignments $w_j^{(i)}$.

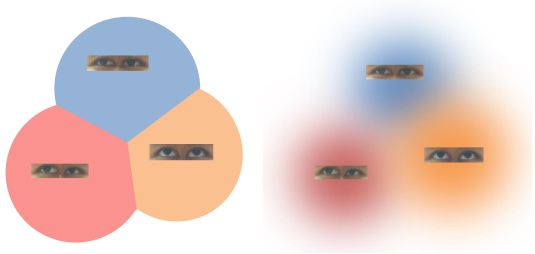


Fig. 2. Diagrams illustrating the clustering methods of both (left) Sugano *et al.* and (right) proposed.

Theoretically, during the Gaussian process regression, we assume a noisy observation model:

$$g_i = f(e_i) + \epsilon_i, \tag{10}$$

where g_i and e_i represents the predicted gaze point and the input eye feature vector respectively, and ϵ_i represents the noise term. Gaussian process regression will basically give g_i as a function of e_i with the noise term $\epsilon_i = \mathcal{N}(0, \varsigma_i^2)$. Noisy observation model assumes that the data (training inputs) has noise modeled by a Gaussian centered around $f(e_i)$. In other words, it assumes that $\overline{e}_i$ from (3) are Gaussian cluster centers, which is correct. Thus by the GMM clustering method, the noisy observation model (10) with the noise term $\epsilon_i = \mathcal{N}(0, \varsigma_i^2)$ is a suitable assumption. This is another justification using GMM as a clustering method.

4. EXPERIMENT

We carry out experiments in two parts. First, we measure the gaze estimation error in degrees after training with a 5 minute video clip which includes four test sequences. The resolution of the video is 1280×760 pixels. Testing is done by the user watching short sequences of moving ground truth points. Next, we train with a long sequence of moving gaze points and test with the same sequence as ground truth. We build 15 dimensional eye features by dividing gray scale eye images into 3×5 cells and averaging them. The methods used in [9, 10] are implemented for comparison in our experiments.

4.1. Calibration-free gaze estimation experiment

A user watches a 5 minute video clip for calibration. A few examples are shown in the 1st column of Fig. 3. For testing, the user watches short sequences of moving gaze points. The moving gaze points are used as ground truth gaze points for testing. The estimation error in degrees is shown in Table 1.

In Fig. 3, the 1st column shows input frames along with eye images and ground truth gaze point (yellow). The 2nd and 3rd columns show resultant probability maps and gaze points (red) from our clustering method and [9, 10] respectively. Our method shows accurate gaze point estimation whereas [9, 10] shows significant deviation.

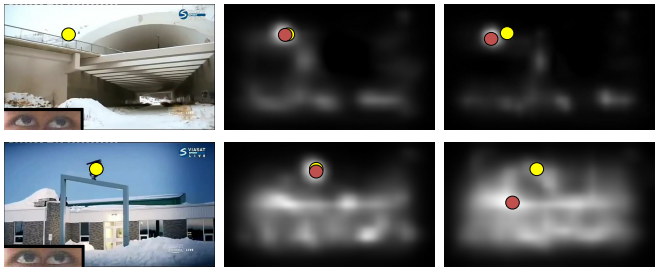


Fig. 3. Eyes and resultant probability maps of both methods.

Test sequence	Sugano *et al.* [9, 10]	Proposed
Test# 1	4.42	2.04
Test# 2	4.25	1.19
Test# 3	5.33	3.16
Test# 4	4.98	3.51
Average	4.78	2.60

Table 1. Calibration-free gaze estimation error (in degrees).

4.2. Gaze trajectory estimation experiment

A user watches a long sequence of moving points. The same sequence is used as ground truth gaze points. This experiment removes all factors that may affect the gaze estimation quality except for the clustering (saliency aggregation) methods. Table 2. shows the estimation error in degrees.

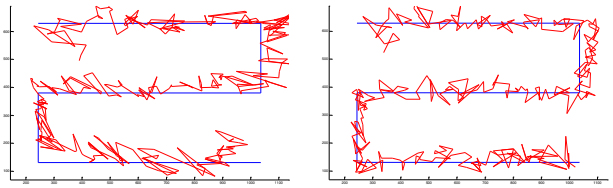
	Sugano *et al.* [9, 10]	Proposed
Error (degrees)	1.58	0.95

Table 2. Gaze trajectory estimation error (in degrees).

Fig. 4 shows the gaze trajectory estimation results. For [9, 10], due to a hard clustering method, the input data holds biased information and thus produces results shown in Fig. 4(a). Notice that it has larger fluctuations and bias along the trajectories. In contrasts, our method reduces fluctuation and bias through a soft clustering method with GMM. This method is also in consensus with the noisy observation model described in (10) which leads to further fluctuation reduction. Due to these improvements we can observe an improved result in Fig. 4(b).

5. CONCLUSION

In this work, we propose a calibration-free gaze estimation by saliency aggregation according to an adaptive GMM-based clustering. This soft clustering method reduces bias and improves the overall estimation quality. In addition, it is in consensus with the noisy observation model of GP. This may be extended to driver safety applications involving front view saliency maps as a future work. For this extension, we must establish a real-time free head pose gaze tracking system.



(a) (b)

Fig. 4. Display of gaze trajectory estimation (red) and ground truth trajectory (blue) results trained by (a) Sugano *et al.* [9, 10] method and (b) proposed method.

Overall, with this theoretical approach, the experimental results show improvement in gaze estimation accuracy.

6. ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No. 2010-0028680).

7. REFERENCES

[1] D.W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32(3), pp. 478–500, 2010.

[2] Kar-Han Tan., D. Kriegman, and N. Ahuja, "Appearance-based eye gaze estimation," in *Proc. IEEE Workshop on Applications of Computer Vision*, 2002, pp. 191–195.

[3] K. Liang, Y. Chahir, M. Molina, C. Tijus, and F. Jouen, "Appearance-based gaze tracking with spectral clustering and semi-supervised gaussian process regression," in *Proc. Conf. on Eye Tracking South Africa*, 2013, pp. 17–23.

[4] J. Chen and Q. Ji, "Probabilistic gaze estimation without active personal calibration," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011.

[5] Y. Sugano, Y. Matsushita, Y. Sato, and H. Koike, "An incremental learning method for unconstrained gaze estimation," in *Proceedings of the 10th European Conference on Computer Vision*, 2008, pp. 656–667.

[6] H. Yamazoe, A. Utsumi, T. Yonezawa, and S. Abe, "Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions," in *Proceedings of the 2008 symposium on Eye tracking research & applications*, 2008, pp. 245–250.

[7] F. Alnajar, T. Gevers, R. Valenti, and S. Ghebreab, "Calibration-free gaze estimation using human gaze patterns," in *Proc. Int. Conf. on Computer Vision*, 2013.

[8] PhiBang Nguyen., J. Fleureau, C. Chamaret, and P. Guillotel, "Calibration-free gaze tracking using particle filter," in *Proc. IEEE Int. Conf. on Multimedia and Expo*, 2013, pp. 15–19.

[9] Y. Sugano, Y. Matsushita, and Y. Sato, "Calibration-free gaze sensing using saliency maps," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010, pp. 2667–2674.

[10] Y. Sugano, Y. Matsushita, and Y. Sato, "Appearance-based gaze estimation using visual saliency," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP(99), 2012.

[11] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proceedings of Advances in neural information processing systems*, 2007, vol. 19, pp. 545–552.